一种融合外部特征的改进主题模型*

杨如意　刘东苏　李　慧
(西安电子科技大学经济与管理学院　西安 710126)

摘要:【目的】在 LDA 模型基础上融合时间和作者特征, 提出动态作者主题(DAT)模型, 更好地揭示文本内容、主题和作者之间的关系。【应用背景】从海量文本中实现特征抽取和语义挖掘已经成为情报研究人员的重要工作。【方法】获取 NIPS 会议论文作为数据集并进行预处理, 按发表年份划分到每个时间片形成一阶马尔科夫链, 使用困惑度确定最优主题数, 并在每个时间片内通过吉布斯采样估算作者主题概率分布和主题词项概率分布。【结果】实验结果表明, 该模型将文档表示为作者主题概率分布和主题词项概率分布, 时间维度上可观测主题强度变化和作者兴趣变化。【结论】DAT 模型能够有效地融合文档内容与外部特征, 实现文本挖掘。

关键词: LDA 模型　DAT 模型　文本挖掘　吉布斯采样
分类号: G202

1　引　言

在当前信息环境下, 文本是最为主要的信息表达方式, 从海量文本中实现特征抽取和语义挖掘已经成为情报研究人员的重要工作。主题模型凭借其在挖掘文本隐含信息的有效性而赢得广泛关注。主题模型从文档生成过程的角度进行建模, 通过统计文档层面的词项共现信息, 抽取出语义上相近的主题, 将文档表示成一组主题, 大幅降低了文档的特征空间维度[1]。2003 年, Blei 等[2]提出 LDA(Latent Dirichlet Allocation)模型, 它是一个三层贝叶斯模型, 将文档看成不同的主题以一定概率分布组成, 每一个主题看成不同的词项以一定概率分布组成。Griffiths 等[3]认为 LDA 模型提取的主题能捕捉到数据中有意义的结构, 从而阐明语义内容, 并对 LDA 模型中的β参数施加 Dirichlet 先验, 使之更加完整。

以 LDA 为代表的主题模型关注文本内容的语义挖掘, 而没有考虑外部特征, 为此本文提出一种基于 LDA 的改进主题模型, 融合了作者和时间两个外部特征, 旨在揭示文档内容、主题和作者之间的动态关系。

2　相关工作

在 LDA 模型之后, 越来越多的研究人员通过扩展主题模型完成文本语义挖掘任务。Blei 等[4]提出 CTM(Correlated Topic Model)模型, 克服了 LDA 模型中不同主题之间弱相关性的缺点, 将主题之间的相关性用一个协方差矩阵表示, 有效地改进了主题抽取的效果。Li 等[5]针对 CTM 只考虑两个主题间关系的不足, 提出了 PAM 模型。其核心思想是用有向无环图(DAG)描述文档中隐含主题之间的结构, 叶子节点是单词, 非叶子节点(主题)可以看成是由所包含的子节点(主题或词项)构成, 那么主题可能是词项概率分布, 也可能是(子)主题概率分布[1]。PAM 模型的缺陷在于, 对主题概率分布进行采样的过程过于复杂, 不易于实现。Rosen-Zvi 等[6]基于 LDA 提出 Author-Topic 模型, 引入文档作者信息, 用于对文档内容和作者的建模, 作者可表示为一组主题的概率分布, 从而发现每个主题下的知名作者。Wang 等[7]向 LDA 模型中添加一个作为观测值的时间随机变量后得到主题随时间变化的主题模型(Topic Over Time, TOT), 认为主题概率分布受到时间的影响, 而时间变量服从 Beta 分布。

通讯作者: 杨如意, ORCID: 0000-0003-4427-9608, E-mail: yangry0801@163.com。
*本文系国家自然科学基金青年基金项目“基于可信语义 wiki 的知识库构建方法与应用研究”(项目编号:71203173)的研究成果之一。

李文波等[8]在 LDA 模型基础上引入文本的类别信息，提出 Labeled-LDA 模型，在各个类别上协同计算隐含主题的分配量，克服了传统 LDA 模型用于分类时强制分配隐含主题的缺陷，有效改进文本分类的性能。王萍[9]验证了使用 LDA 主题模型进行文献知识挖掘的可行性，对文献的文本信息和作者信息进行联合建模，提出多维度文献知识挖掘方法。胡吉明等[10]基于动态 LDA 模型进行主题演化与挖掘，在每个时间片内采用 LDA 模型进行主题挖掘，其不足之处在于本质上还是对文本内容进行挖掘，缺少外部特征。

针对主题模型仅限于分析文档的内部特征而不考虑外部特征的缺陷，本文改进思路是在LDA模型的基础上融合文本内外部特征。借鉴胡吉明等[10]将文本按时间划分的思想，但时间片之间的状态依赖关系不同，创新点是在主题采样的过程中加入先验参数——作者主题概率分布，通过作者主题概率分布发现作者研究兴趣变化。

3 基于 LDA 的改进主题模型

动态主题模型[11](Dynamic Topic Model)考虑了时间和文本主题的连续性，作者主题模型[6](Author Topic Model)考虑了作者和文本主题之间的关系，两者都是基于 LDA 模型引入了一种外部特征。本文结合上述两个模型的优势，提出动态作者主题模型(Dynamic Author Topic, DAT)，首先将文档集划分到不同的时间片内，在每个时间片内对子文档集进行建模分析，文档中可观测变量是作者和词项，每个作者都对应一个在主题上的多项分布，每个主题都对应一个在词项上的多项分布，如图 1 所示：

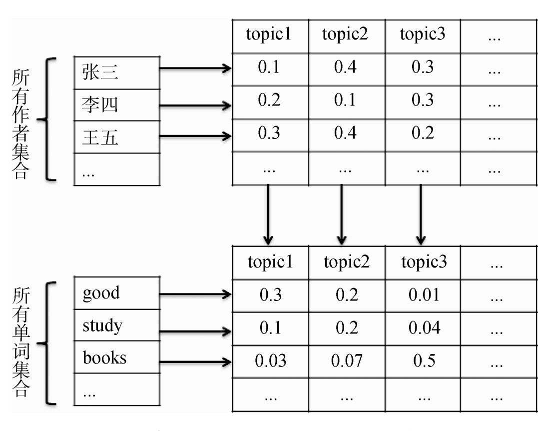


图 1 作者主题词项的概率分布示意图

基于这个思想，文档可表示为作者主题概率分布和主题词项概率分布，在时间维度上进行观测，还可发现作者研究兴趣变化、主题内容和强度变化。下面从模型输入、基本假设、模型表示和参数估计这 4 个方面对 DAT 模型进行论述。

3.1 模型输入和基本假设

主题模型的主要输入是文档集合，Griffiths 等[3]将文本中的词分为两大功能：一个是语义功能，用于表示文档主题，也就是特征词；另一个是语法功能，这些词的存在是为了让整个句子的生成过程看起来更像一个整体或者说更符合语言规范，比如虚词、代词和量词等。这些高重复性的非特征词和文档主题无关，需要在预处理中进行停用词(Stop Words)去除。经过预处理后的文档集，实质上就是文档集的特征词序列。

另外一个重要的输入是主题个数 T，通常 T 的大小需要在模型训练前指定，而且存在一定的经验性，确定最优 T 的简单方法是用不同的 T 值进行重复实验，也可采用困惑度(Perplexity)[12]指标确定最优主题数。

DAT 模型包含的基本假设主要有：文档的词顺序是可交换的；文章各个主题之间不相关或者弱相关；作者顺序是可交换的，即文档中每个词均匀地随机地由某个作者产生；不同时间片的模型参数满足一阶马尔科夫假设，即仅与前一时间片的模型参数有关。

3.2 模型表示

为了清晰地阐述 DAT 模型，对本文所使用的符号进行说明，如表 1 所示：

表 1 符号说明

符号	描述
D	文档的数量
T	文档集所有主题的数量
V	文档集所有词项的数量
A	文档集所有作者的数量
N_d	文档 d 中特征词的数量
A_d	撰写文档 d 的作者数量
a_d	撰写文档 d 的作者向量
θ_x	作者 x 的主题概率分布
ϕ_t	主题 t 的词项概率分布
x	文档 d 中采样的某个作者
$z_{d,n}$	文档 d 中第 n 个单词的主题分配
$w_{d,n}$	文档 d 中第 n 个词项
α	θ的 Dirichlet 先验参数
β	ϕ的 Dirichlet 先验参数

DAT 模型的概率图表示如图 2 所示，实线箭头表示变量之间的条件依赖关系，虚线箭头表示不同时间片内的参数渐变，通过参数α和β的渐变构建不同时间片的文档子集之间的状态依赖，矩形表示重复采样(生成)，其右下角字母表示采样次数，可观察变量是文档的作者和词项，表示为填充阴影的圆。该模型中，在每个时间片内文档子集 D 的产生过程如下：

(1) 对于每个主题 $t\in[1, T]$，采样ϕ_t~Dirichlet(β)；

(2) 对于每个作者 $x\in[1, A]$，采样θ_x~Dirichlet(α)；

(3) 对于每篇文档中的每个词项 $w_{d,n}$：

① a 采样一个作者 $x_{d,n}$~Uniform(a_d)；

② b 采样一个主题 $z_{d,n}$~Multinomial($\theta x_{d,n}$)；

③ c 采样一个词项 $w_{d,n}$~Multinomial($\phi z_{d,n}$)；

(4) 重复步骤(3)的采样过程 N_d 次，生成文档 d 的全部特征词；

(5) 重复步骤(4)的采样过程 D 次，生成整个文档子集。

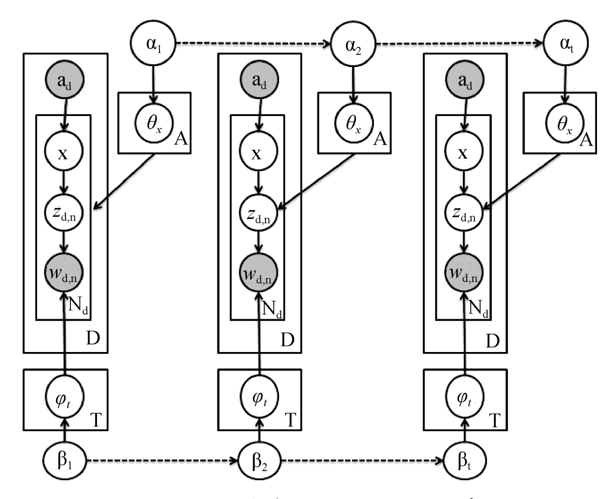


图 2　动态作者主题模型的图表示

3.3　参数估计

对 LDA 模型进行参数估计的方法有很多，常用的有 VB (Variational Bayesian Inference)算法[2]，EP (Expectation-Propagation) 算法[13]，Collapsed Gibbs Sampling[14]等。本文选择 Gibbs 抽样方法，它是一种快速高效的 MCMC (Markov Chain Monte Carlo)抽样方法，利用每个变量的条件概率分布实现从联合分布中抽样，通过反复抽样迭代，得到参数估计值。

在 LDA 模型中有两组待估计参数：文档主题概率分布和主题词项概率分布，在 DAT 模型中，需要估计的也是两组参数：作者主题概率分布θ和主题词项概率分布ϕ。

对于每个时间片内的文档子集，通过 Gibbs 采样为每个词项分配了主题 z 和作者 x，利用 Dirichlet 分布的期望，推导如下计算公式：

$$\phi_{t,v}=\frac{n_t^{(v)}+\beta}{\sum_{v=1}^{V}n_t^{(v)}+V\beta} \tag{1}$$

$$\theta_{a,t}=\frac{n_a^{(t)}+\alpha}{\sum_{t=1}^{T}n_a^{(t)}+T\alpha} \tag{2}$$

其中，$\phi_{t,v}$ 表示主题 t 中包含词项 v 的概率，$n_t^{(v)}$ 表示词项 v 分配到主题 t 的次数，$\theta_{a,t}$ 表示作者 a 包含主题 t 的概率(a 对 t 感兴趣的概率)，$n_a^{(t)}$ 表示主题 t 分配到作者 a 的次数。

在典型的语言建模应用中，Dirichlet 分布经常被用来刻画词项分布的不确定性[3]，本文中ϕ_t 服从参数为β的 Dirichlet 分布，θ_x 服从参数为α的 Dirichlet 分布。图 2 中虚线箭头表示相邻时间片的超参数渐变，依据 3.1 节中的模型假设，本文用 u，v 分别刻画α和β的渐变权重，定义如下：

$$\beta_t=u_t\times\beta_{t-1} \tag{3}$$

$$u_t=(\text{Token})_t/(\text{Token})_{t-1} \tag{4}$$

$$\alpha_t=v_t\times\alpha_{t-1} \tag{5}$$

$$v_t=(\text{Author})_t/(\text{Author})_{t-1} \tag{6}$$

Token 是当前时间片内文档子集的词项总数，Author 是当前时间片内文档子集的作者总数，因此当第一个时间片内的超参数取值确定时，其后的取值均可确定。

3.4　模型对比

从推断方法、时间、作者三个方面，对比 4 个扩展主题模型的区别，如表 2 所示：

表 2　模型对比

比较项	作者主题模型	动态主题模型	作者主题演化模型	动态作者主题模型
推断方法	Gibbs 采样	变分期望最大化	Gibbs 采样	Gibbs 采样
如何处理时间	/	离散，一阶马尔科夫	连续，Beta 分布	离散，一阶马尔科夫
是否包含作者	是	/	是	是

4　实验过程

实验过程设计如图 3 所示。实验的机器是 HP

ProDesk 600 台式电脑，CPU 是 Intel i5-4590 处理器，内存 4GB，系统是 Windows7 64bit 版本。使用 Eclipse 开发工具，用 Java 语言编写程序完成数据抽取、分词等预处理工作并实现 DAT 模型，最终得到文档集的主题词项概率分布和作者主题概率分布。

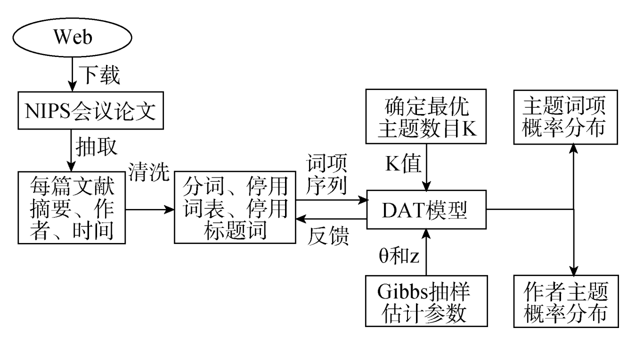


图 3　实验过程

4.1　数据集和文本预处理

选取 NIPS(Neural Information Processing Systems) 会议论文作为实验数据。在 Web of Science 核心集数据库中，以"NIPS"为会议关键词进行检索，时间范围是 1997-2001 年，得到 758 条记录，检索结果的出版年份分布如图 4 所示，将每个年度的检索结果集导出为纯文本格式，得到原始实验数据集。

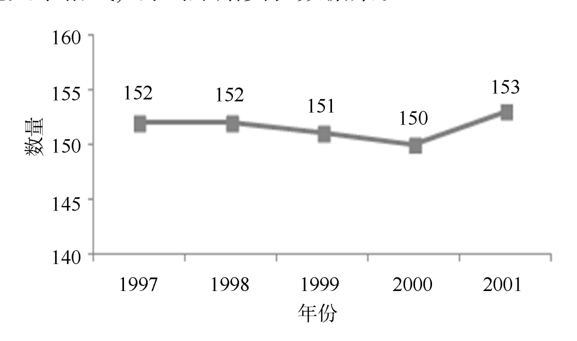


图 4　数据集的时间分布

对以上文本数据进行预处理，本文只对论文中的摘要、出版年份、作者进行分析，不保留其他特征。利用正则表达式去除虚词、代词、量词等词项，并去除词频低于 3 的词项，按照年份进行分类汇总，得到 DAT 模型的输入数据集。

4.2　确定主题数

主题模型中两个超参数的经验值取值一般为 $\alpha=50/T$，$\beta=0.01$[15]，对第一个时间片内的超参数取值也做上述处理。主题数目通常由用户输入，其取值对于模型中的主题抽取和拟合性能影响较大，其最佳值的确定主要通过两种方式：词汇被选概率和困惑度[2]。困惑度是从模型泛化能力衡量主题模型对于新文本的预测能力，困惑度越小表示模型的泛化能力越强，产生文档的性能越高，能够较为全面地评价模型效果。本文选取困惑度作为评测指标，Blei 等[2]定义一个数据集的主题模型的困惑度为：

$$perplexity(w_d \mid a_d) = \exp\left[-\frac{\ln p(w_d \mid a_d)}{N_d}\right]$$

其中，w_d 是文档 d 的特征词向量，表示所有词项，a_d 是文档 d 的作者向量，表示所有作者。$p(w_d \mid a_d)$ 表示在给定一组作者的情况下生成特征词向量的概率，Rosen-Zvi 等[6]在作者主题模型中给出了其推导算式如下：

$$p(w_d \mid a_d) = \int d\theta \int d\phi p(\theta \mid D^{train}) p(\phi \mid D^{train}) \times \prod_{m=1}^{M_d}\left[\frac{1}{A_d}\sum_{i\in a_{d,j}} \theta_{ij}\phi w_{mj}\right]$$

对于不同的 K 取值，分别进行 Gibbs 抽样，迭代次数 500，困惑度取值的变化情况如图 5 所示。因此，选取的最优主题数目是 T=50。

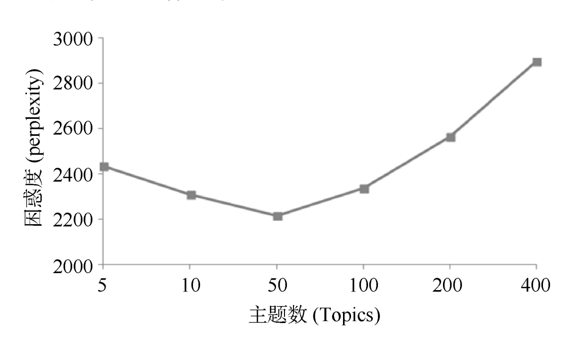


图 5　困惑度随主题数的变化

4.3　结果分析

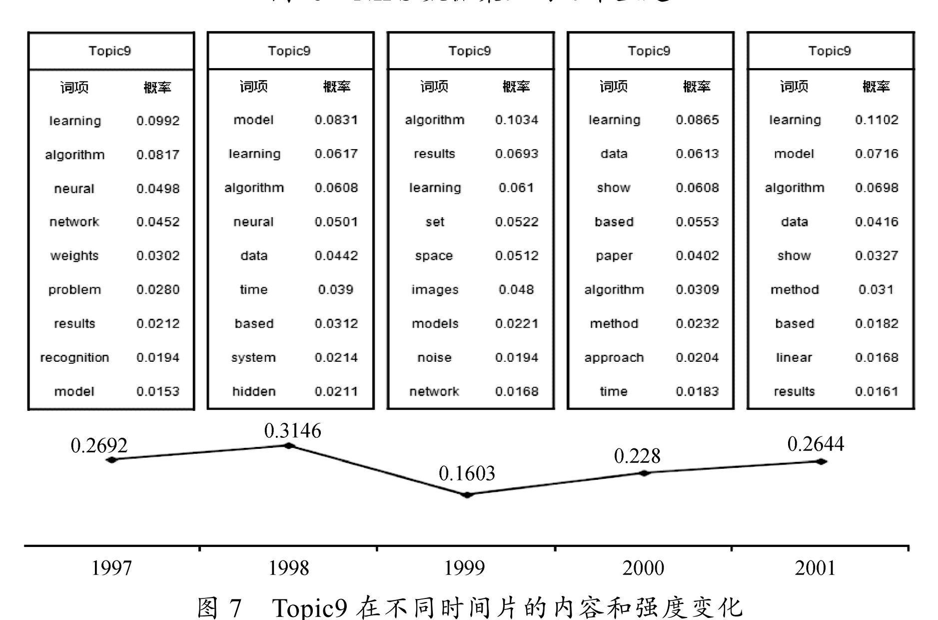
利用 DAT 模型，对实验数据进行处理后得到作者主题分布和主题词项分布，图 6 给出了 1997 年文档集的部分主题的表示，每个主题表示为概率最大的前 10 个词项和前 10 个作者。

实验结果显示，主题 9 关注"神经学习算法"方面的内容，相关的研究人员有 Hinton GE, Dayan P 等人；主题 11 关注"模型应用"，相关研究人员有 Sejnowski TJ, Graepel T 等人；主题 20 关注"数据分类和识别"，相关研究人员有 Singh S, Seung HS 等人；主题 29 关注"似然估计"，相关研究人员有 Bishop C, Koch C 等人。通过 DAT 主题模型对文档集的隐含主题进行抽取，将隐含主题表示为作者和词项的概率分布。观测同一个

主题在不同时间片内的词项概率变化，可实现主题演化分析。图 7 揭示了主题 9 在每个时间片内文档子集的词项分布变化，下方的曲线表示主题的概率强度变化。从作者主题分布中，可发现作者在各个时间片对同一主题的兴趣强弱变化。图 8 揭示了作者 Willams C 对 Topic20 的研究兴趣变化。

Topic9		Topic11		Topic20		Topic29	
词项	概率	词项	概率	词项	概率	词项	概率
learning	0.0992	application	0.0516	classification	0.1026	likelihood	0.0683
algorithm	0.0817	models	0.0462	data	0.0559	states	0.0445
neural	0.0498	approximation	0.0343	recognition	0.0371	training	0.0378
network	0.0452	estimating	0.0312	method	0.0323	similar	0.0265
weights	0.0302	relative	0.0287	parameters	0.0202	infinite	0.0243
problem	0.0280	grid	0.0217	function	0.0137	individual	0.0204
results	0.0212	regularize	0.0181	number	0.0120	bayesian	0.0180
recognition	0.0194	speech	0.0127	linear	0.0116	estimation	0.0168
model	0.0168	tracking	0.0115	relative	0.0112	experience	0.0164
data	0.0160	stages	0.0108	derive	0.0097	based	0.0132
作者	概率	作者	概率	作者	概率	作者	概率
Hinton GE	0.1184	Sejnowski TJ	0.0845	Singh S	0.2017	Bishop C	0.1059
Dayan P	0.0469	Graepel T	0.0631	Seung HS	0.0534	Koch C	0.0872
Scholkopf B	0.0357	Freeman WT	0.0604	Williams C	0.0439	Mozer M	0.0431
Vapnik V	0.0345	Winther O	0.0601	Tishby N	0.0427	Lee DD	0.0293
Singer Y	0.0285	Ruppin E	0.0578	Smola A	0.0318	Jordan M	0.0287
Tresp V	0.0282	Smola AJ	0.0354	Saad D	0.0116	Meir R	0.0176
Koller D	0.0176	Movellan J	0.0312	Opper M	0.0109	Horn D	0.0173
Sollich P	0.0163	Barto AG	0.0308	Muller KR	0.0109	Sutton R	0.0169
Cole R	0.0160	Becker S	0.0297	Ratsch G	0.0097	Tipping M	0.0068
Hancock R	0.0151	Attias H	0.0263	Bengio Y	0.0092	Platt J	0.0064

图 6　NIPS 数据集上的 4 个主题



Topic9		Topic9		Topic9		Topic9		Topic9	
词项	概率	词项	概率	词项	概率	词项	概率	词项	概率
learning	0.0992	model	0.0831	algorithm	0.1034	learning	0.0865	learning	0.1102
algorithm	0.0817	learning	0.0617	results	0.0693	data	0.0613	model	0.0716
neural	0.0498	algorithm	0.0608	learning	0.061	show	0.0608	algorithm	0.0698
network	0.0452	neural	0.0501	set	0.0522	based	0.0553	data	0.0416
weights	0.0302	data	0.0442	space	0.0512	paper	0.0402	show	0.0327
problem	0.0280	time	0.039	images	0.048	algorithm	0.0309	method	0.031
results	0.0212	based	0.0312	models	0.0221	method	0.0232	based	0.0182
recognition	0.0194	system	0.0214	noise	0.0194	approach	0.0204	linear	0.0168
model	0.0153	hidden	0.0211	network	0.0168	time	0.0183	results	0.0161

图 7　Topic9 在不同时间片的内容和强度变化

0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0.0439
0.0151
0.0264
0.0612
0.0207
1997 1998 1999 2000 2001

图 8　作者 Willams C 对 Topic20 的兴趣变化

5　结　语

从海量科技文献中自动挖掘隐含主题、作者研究兴趣及其变化，是情报研究的重要内容之一。目前以 LDA 为代表的主题模型对文本内容的特征抽取得到

了广泛应用，但是缺少对多个外部特征的融合分析。本文在研究动态主题模型和作者主题模型的优势后，引入时间和作者两个外部特征进行扩展，构造动态作者主题模型，将文档表示为作者主题概率分布和主题词项概率分布，并以 NIPS 会议的论文作为实验数据集，通过吉布斯采样估算参数，验证了模型的有效性。本文的不足之处在于，模型的隐含假设是文档作者服从均匀分布，与作者的排序无关，文档主题服从多项分布，不同主题之间具备弱相关性，这与实际语料不符。

本文在以下两个方面值得进一步研究。

(1) 通过研究不同时间片内的主题词项概率分布，发现主题内容变化和强度变化。主题内容变化可表示为主题的词项概率分布随着时间变化增大或减小，带来主题语义上的变迁；主题强度变化可表示为文档集中同一个主题在不同时间片上的概率大小，带来文档主题的变迁，从而实现主题演化分析。

(2) 研究作者主题概率分布，研究人员在不同时间片内对同一个主题的关注度有强弱之分，表现为研究兴趣的变迁，从而发现作者的研究兴趣变化。

参考文献：

[1] 徐戈，王厚峰.自然语言处理中主题模型的发展[J]. 计算机学报，2011, 34(8): 1423-1436. (Xu Ge, Wang Houfeng. The Development of Topic in Natural Language Processing [J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436.)

[2] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[3] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academy of Sciences, 2004, 101(S1): 5228-5235.

[4] Blei D M, Lafferty J D. A Correlated Topic Model of Science [J]. The Annals of Applied Statistics, 2007, 1(1): 17-35.

[5] Li W, McCallum A. Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations [C]. In: Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 557-584.

[6] Rosen-Zvi M, Griffths T, Steyvers M, et al. The Author Topic Model for Authors and Documents [C]. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2004: 487-494.

[7] Wang X, McCallum A. Topic Over Time: A Non-Markov Continuous-Time Model of Topical Trends [C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM, 2006: 424-433.

[8] 李文波，孙乐，张大鲲. 基于 Labeled-LDA 模型的文本分类新算法[J].计算机学报，2008, 31(4): 620-627. (Li Wenbo, Sun Le, Zhang Dakun. Text Classification Based on Labeled-LDA Model [J]. Chinese Journal of Computers, 2008, 31(4): 620-627.)

[9] 王萍.基于概率主题模型的文献知识挖掘[J]. 情报学报，2011, 30(6): 583-590. (Wang Ping. Literature Knowledge Mining Based on Probabilistic Topic Model [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(6): 583-590.)

[10] 胡吉明，陈果. 基于动态 LDA 主题模型的内容主题挖掘与演化[J]. 图书情报工作，2014, 58(2): 138-142. (Hu Jiming, Chen Guo. Mining and Evolution of Content Topics Based on Dynamic LDA [J]. Library and Information Service, 2014, 58(2): 138-142.)

[11] Blei D M, Lafferty J D. Dynamic Topic Models [C]. In: Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 113-120.

[12] Azzopardi L, Girolami M, Van Risjbergen K, et al. Investigating the Relationship Between Language Model Perplexity and IR Precision-Recall Measure [C]. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.New York: ACM Press, 2003: 369-370.

[13] Minka T, Lafferty J. Expectation Propagation for the Generative Aspect Model [C]. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence.Morgan Kaufmann Publishers Inc., 2002: 352-359.

[14] Teh Y W, Newman D, Welling M. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation [C]. In: Proceedings of the Neural Information Processing Systems Conference.2006.

[15] 史庆伟，乔晓东，徐硕，等.作者主题演化模型及其在研究兴趣演化分析中的应用[J].情报学报，2013, 32(9): 912-919. (Shi Qingwei, Qiao Xiaodong, Xu Shuo, et al. Author-Topic Evolution Model and Its Application in Analysis of Research Interests Evolution [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(9): 912-919.)

作者贡献声明：

杨如意，刘东苏：提出研究思路改进模型，设计研究方案；
杨如意：采集、清洗、分析数据，进行实验；
杨如意，李慧：分析结果，起草论文；
杨如意，刘东苏：论文最终版本修订。

收稿日期: 2015-07-17
收修改稿日期: 2015-08-30

An Improved Topic Model Integrating Extra-Features

Yang Ruyi Liu Dongsu Li Hui
(School of Economics and Management, Xidian University, Xi'an 710126, China)

Abstract: [**Objective**] In order to reveal the relationships between contents, topics and authors of documents, this paper presents the Dynamic Author Topic (DAT) model which extends LDA model. [**Context**] Extracting features from large-scale texts is an important job for informatics researchers. [**Methods**] Firstly, collect the NIPS conference papers as data set and make preprocessing with them. Then divide data set into parts by published time, which forms a first-order Markov-chain. Then use perplexity to ensure the number of topics. At last, use Gibbs sampling to estimate the author-topic and topic-words distributions in each time slice. [**Results**] The results of experiments show that the document is represented as probability distributions of topics-words and authors-topics. On the dimension of time, the revolution of authors and topics can be observed. [**Conclusions**] DAT model can integrate contents and extra-features efficiently and accomplish text mining.

Keywords: LDA model DAT model Text mining Gibbs sampling

美国德州农工大学图书馆宣布加入 Kuali OLE

美国德州农工大学图书馆于近日宣布计划加入Kuali OLE(Open Library Environment, 开放图书馆环境)和Kuali基金会, 双方正式成为合作伙伴。Kuali OLE是一个企业规模的、基于云计算的、源于科研社区的图书馆管理系统, 其建立在开放标准上, 目的是构建一个健壮的企业工作流引擎, 为图书馆业务流程的高效管理提供保障。OLE支持多种学术信息资源和格式, 并且正在兴建的过程中, 由学术界和研究图书馆社区负责管理和运行。

"我们很高兴加入Kuali OLE, 这是一个由高校联盟开发并且服务于高校的项目," 德州农工大学图书馆数字图书馆倡议工程负责人Michael Bolton表示, "这种伙伴关系有助于我们以高效和有效的方式开发专门适应于研究图书馆的软件。"

参与Kuali OLE的机构采取结构化的方法开发软件, 旨在取代现有的图书馆管理系统。德州农工大学图书馆的加入将会进一步确保他们在早期测试和开发上的优势。这也使得德州农工大学图书馆在决定该系统需要解决哪些问题时有一定的话语权, 比如整合图书馆资源和电子资源。

"开源的OLE系统是德州农工大学下一代图书馆管理系统的正确选择," 德州农工大学图书馆馆长David Carlson说, "我们渴望加入一个有着众多其他研究图书馆的领导联盟, 为开发下一代的开源软件付出努力。"

芝加哥大学、利哈伊大学、伦敦大学亚非学院目前正在使用 Kuali OLE 系统, 杜克大学也计划在 2016 年使用该系统。德州农工大学计划在 2017 年开始使用该系统。

(编译自: http://library.tamu.edu/news/2016/01/KualiOLE_Partnership.html)

(本刊讯)